° 30 April 2001

Article reference:
CB20.300401
Coffee Break archives

nature REVIEWS

Story by Carina Dennis, Nature Reviews Genetics

# fadA rpoB

#### Gene distribution map. The locations of *M. leprae* protein coding genes are represented on this map.

Click on the figure for more information.

# Honey, I shrunk the genome

""Getting by with the bare minimum seems to be the modus operandi of Mycobacterium leprae the causal agent of leprosy. Its genome sequence reveals that it has undergone massive genome 'downsizing' over time, discarding more than half its genes and rendering it the most striking example of genome reduction in a microbial pathogen.

""The leprosy bacillus is famed for being the first microorganism definitively shown to be associated with human disease. It evades the host's immune response by invading and propagating inside the vacuoles of macrophages called phagosomes. From there, it infects the Schwann cells of the peripheral nervous system, where it disrupts myelin production, thus leading to the characteristic features of leprosy, which include skin lesions and sensory loss.

efforts to grow it in the laboratory. This is due in part to its excruciatingly slow growth it divides only once every two weeks, slower than any other bacterium and its requirement for an elusive cocktail of nutrients. But the sequencing of its genome by Stewart Cole and colleagues now uncovers clues to its unique metabolism and unusual 'reductive evolution'.

"""M. leprae seems to have completely dispensed with or substantively reduced certain metabolic pathways, including oxidative and anaerobic respiratory chains. The enzymes for breaking down host-derived lipids, a means by which many mycobacterial pathogens derive their energy, are also drastically reduced in M. leprae. By contrast, most anabolic pathways seem to be intact, indicating that M. leprae depends on these pathways to survive in the nutrient-poor microenvironment of phagosomes.

"""Comparison of the genome sequence of *M. leprae* with that of its cousin *Mycobacterium tuberculosis* indicates that the former has undergone substantial downsizing, losing more than 2,000 genes since its divergence from a common mycobacterial ancestor. Presumably its genes became inactivated once their functions were no longer essential to survival, and this was followed by genome shrinkage through rearrangements and deletions.

""The availability of the genome sequence will aid efforts to

# Live PubMed searches

| Leprosy bacillus genome | Tuberculous & leprotic metabolism | REVIEWS |

#### TaxPlot tutorial

See protein homologs in complete genomes context

# Further NCBI resources

Taxonomic distribution of protein homologs M. leprae genome via ftp M. leprae complete genome COG table

define virulence factors of the leprosy bacillus, of which few are presently known, and will provide inroads for developing new vaccines and diagnostic tests. Also, comparative genomic analyses with *M. tuberculosis* are likely to yield insight into the pathogenicity of the causal agent of tuberculosis and into potential drug targets based on the set of genes crucial to the survival of both organisms.

Comments? \*Questions?
We would welcome feedback on NCBI's Coffee Break.
Email to: info@ncbi.nlm.nih.gov

Download this article in Adobe Acrobat PDF format



#### Coffee Break

- Article
- TaxPlot tutorial
- View analysis

#### PubMed Links

- Overview
- Search PubMed

#### **BLAST Links**

- Overview
- BLAST page
- Statistics course

## Honey, I shrunk the genome

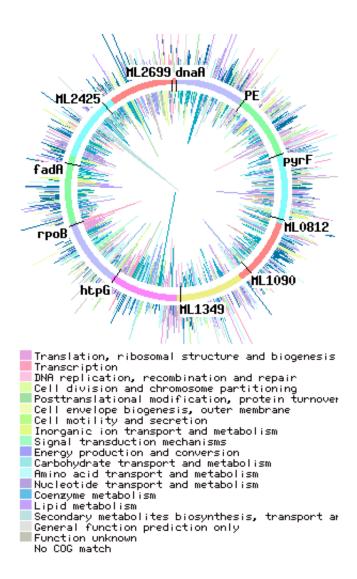


Figure 1. Protein coding genes distribution map for Mycobacterium leprae. The leprosy bacillus genome contains numerous examples of gene deletion and decay. The relative locations of various genes in the genome are depicted in the map above. Protein coding genes are color coded in the map according to their classification within clusters of orthologous groups (COGs) functional categories. COGs represent proteins or groups of paralogs that are found in at least 3 phylogenetically-distant genomes. For more information about COGs, see Science 1997

Oct 24:278(5338):631-7.

# Coffee Break

**BLAST** PubMed Nucleotide Protein **PopSet** Taxonomy Help Genome Structure

#### Mycobacterium leprae strain TN complete genome """ Microbial genomes

Feature table: **BLAST** protein homologs:

Total Bases: 3268203 bp Protein coding genes: 2720

COGs°(Clusters of Orthologous Groups) Structural RNAs: 50 Completed: Feb 20, 2001. Taxa° (Sequences grouped by superkingdom)

Contributor: Sanger Centre "See genome at Sanger Centre Download chromosome sequence data from NCBI FTP site

Accession: NC 002677

Examining the differences between protein sequences of various organisms gives insight into the origin of genes and the relationship between species. TaxPlot provides a means to compare the similarity of query genome proteins to different species.

To compare the genome proteins of *Mycobacterium leprae* with those from other organisms, click on the link marked by the red arrow below. (Note: selecting other links will take you out of this tutorial.)

**NEW** BLAST your query sequence against complete genome protein database

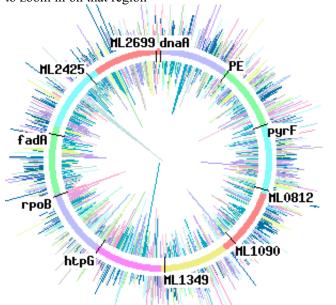
NEW See Conserved Domain Database search - precomputed results

See protein homologs in complete genomes context - TaxPlot

Find Go Start from : Search for gene

#### Protein coding genes distribution map

To see map locations of genes, click on a region in the map, to zoom in on that region



#### Gene Classification based on COG functional categories

Translation, ribosomal structure and biogenesis Transcription DNA replication, recombination and repair Cell division and chromosome partitioning Posttranslational modification, protein turnover Cell envelope biogenesis, outer membrane Cell motility and secretion Inorganic ion transport and metabolism Signāl transduction mechanisms Energy production and conversion Carbóhydrate transport and metabolism Amino acid transport and metabolism Nucleotide transport and metabolism Coenzyme metabolism Lipid metabolism Secondary metabolites biosynthesis, transport an General function prediction only Function unknown No COG match

Organism: Mycobacterium leprae

Genetic Code: 11

Lineage: Bacteria; Firmicutes; Actinobacteria; Actinobacteridae; Actinomycetales;

Corynebacterineae; Mycobacteriaceae; Mycobacterium.

#### Massive gene decay in the leprosy bacillus

Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Ganier,T., Churcher,C., Harris,D., Mungall,K., Basham,D., Brown,D., Chillingworth,T., Connor,R., Davies,R.M., Devlin,K., Duthoy,S., Feltwell,T., Fraser,A., Hamlin,N., Holroyd,S., Hornsby,T., Jagels,K., Lacroix,C., Maclean,J., Moule,S., Murphy,L., Oliver, Quail,M.A., Rajandream,M.-A., Rutherford,K.M., Rutter,S., Seeger,K., Simon,S., Simmonds,M., Skelton,J., Squares,R., Squares,S., Stevens,K., Taylor,K., Whitehead,S., Woodward,J.R. and Barrell,B.G. Nature 409, 1007-1011 (2001)

#### **Direct Submission**

Parkhill,J.

Submitted (20-FEB-2001) Submitted on behalf of the Mycobacterium leprae sequencing teams, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK Unitie de Genetique Moleculaire Bacterienne, Institut Pasteur, 28 rue du Docteur Rouux, 75724, Paris Cedex, France. E-mail: parkhill@sanger.ac.uk

**COMMENT** "Notes: Details of M. leprae sequencing at the Sanger Centre are available from http://www.sanger.ac.uk/Projects/M\_leprae/ A relational datbase containing the M. leprae sequences is available from http://genolist.pasteur.fr/Leproma/.

CONTIG "join (AL583917: 1...344050, AL583918: 51...344050,

AL583919:51..312050, AL583920:51..348450, AL583921:51..332450,

AL583922:51..318200, AL583923:51..348950, AL583924:51..342300,

AL583925:51..308950,AL583926:51..269203)

Comments and suggestions to: [info@ncbi.nlm.nih.gov]



**BLAST** 

PubMed

Nucleotide

Protein

Genome

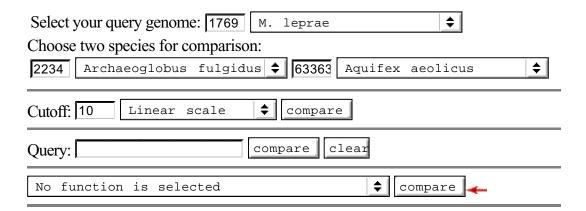
Structure

Taxonomy

#### Protein homologs in Complete Microbial / <u>Eukaryotic</u> genomes

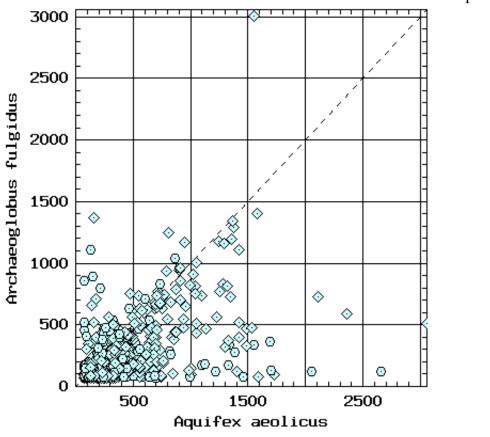
The initial Taxplot view shows the genome proteins of *M. leprae* plotted against two default organisms. Scroll down to view the chart.

Highlight the proteins involved in translation by selecting **Translation**, **ribosomal structure and biogenesis** from the last pull-down menu. Then click on the **compare** button to the immediate right of the menu (marked by the red arrow).



Distribution of Mycobacterium leprae homologs

160 hits 4 equal hits



356 hits

520 total hits Each circle represents a single query genome protein, plotted by its BLAST scores to the highest scoring protein from each of the selected organisms. Click on the protein(s) of interest or enter a query string to see the homologs in two chosen organisms. See HELP for more details.

Click on the protein(s) of interest or enter a query string to see the homologs in two chosen organisms



**BLAST** 

PubMed

Nucleotide

Protein

Genome

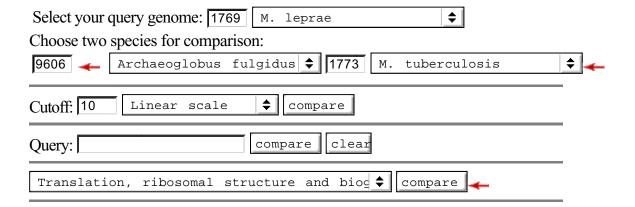
Structure

Taxonomy

#### Protein homologs in Complete Microbial / <u>Eukaryotic</u> genomes

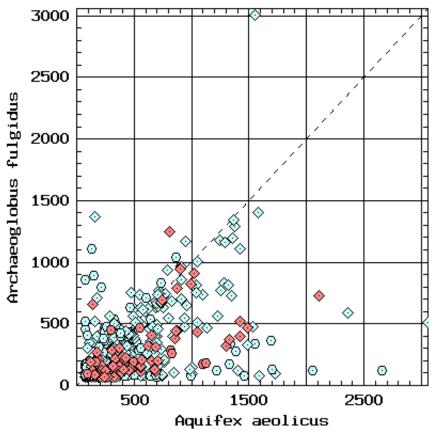
Proteins involved in translation are now highlighted in red. Proteins to the left of the diagonal line are more closely related to *Archaeoglobus fulgidus*, while those that fall to the right of the diagonal are more closely related to *Aquifex aeolicus*.

*M. leprae* genome proteins may be compared to those from other organisms. TaxID 9606 (*H. sapiens*) has been entered manually on the left, and *M. tuberculosis* has been selected in the drop-down menu on the right. Click on the **compare** button marked by the red arrow to continue.



Distribution of Mycobacterium leprae homologs

160 hits 4 equal hits



356 hits

#### 520 total hits, from which 72 are selected:

Each circle represents a single query genome protein, plotted by its BLAST scores to the highest scoring protein from each of the selected organisms. Click on the protein(s) of interest or enter a query string to see the homologs in two chosen organisms. See HELP for more details.

	Definition	Blast2Seq	GenBank Blink
	leucyl-tRNA synthase [Mycobacterium leprae]		<u>13092440</u> =>
aa	valyl-tRNA synthetase (valS) [Archaeoglo	<u>172</u>	<u>11499806</u> =>
	prolyl tRNA synthetase [Mycobacterium leprae]		<u>13093368</u> =>
aa	prolyl-tRNA synthetase (proS) [Archaeogl	<u>656</u>	<u>11499201</u> =>
	proline-tRNA synthetase - Aquifex aeolic	<u>141</u>	<u>7437710</u> =>

Note: this table has been truncated in order to conserve space

**BLAST** 

PubMed

Nucleotide

Protein

Genome

Structure

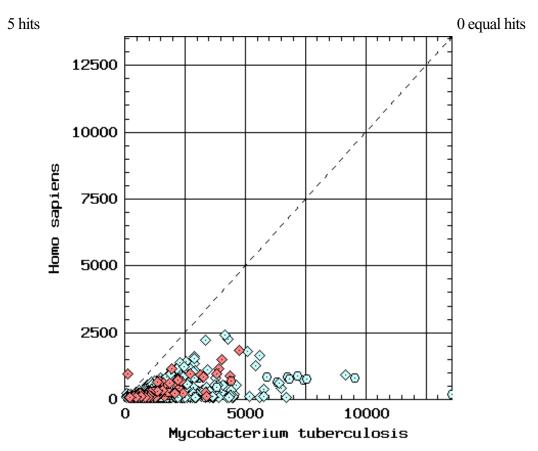
Taxonomy

#### Protein homologs in Complete Microbial / Eukaryotic genomes

As expected, the majority of *M. leprae* genome proteins are more closely related to *M. tuberculosis* proteins than to *H. sapiens* proteins. Click <u>HERE</u> to view the final analysis.

Select your query genome: 1769 M. leprae
Choose two species for comparison:
9606 Homo sapiens 💠 1773 M. tuberculosis 💠
Cutoff: 10 Linear scale 💠 compare
Query: clear
Translation, ribosomal structure and biog \$ compare

#### Distribution of Mycobacterium leprae homologs



#### 928 total hits, from which 103 are selected:

Each circle represents a single query genome protein, plotted by its BLAST scores to the highest scoring protein from each of the selected organisms. Click on the protein(s) of interest or enter a query string to see the homologs in two chosen organisms. See HELP for more details.

	Definition	Blast2Seq	GenBank Blink
	leucyl-tRNA synthase [Mycobacterium leprae]		13092440 =>
aa	leucyl-tRNA synthetase, mitochondrial; K	<u>863</u>	<u>7661872</u> =>
	LEUCYL-TRNA SYNTHETASE (LEUCINETRNA LI	4369	<u>2501027</u> =>
	prolyl tRNA synthetase [Mycobacterium leprae]		<u>13093368</u> =>
aa		<u>973</u>	<u>12722495</u> =>
	PROLYL-TRNA SYNTHETASE (PROLINETRNA LI	124	<u>2501040</u> =>
aa	50S RIBOSOMAL PROTEIN L34	<u>224</u>	<u>1710543</u> =>

Note: this table has been truncated in order to conserve space

### Honey, I shrunk the genome

Mycobacterium leprae protein homologs in complete genomes context

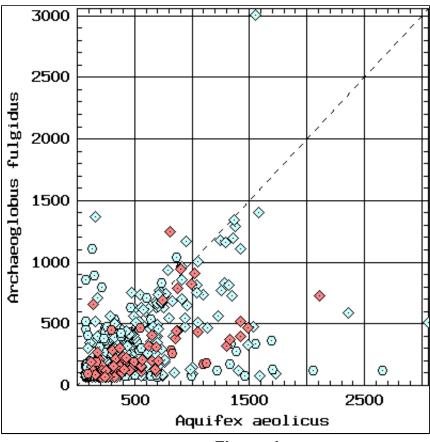


Figure 1

The analysis of protein similarities between organisms gives insight into their evolutionary relationships. Figure 1 shows the TaxPlot of *Mycobacterium leprae* genome (NC\_002677) proteins compared to proteins from the genomes of *Archaeoglobus fulgidus* and *Aquifex aeolicus*. Each marker on the graph is derived from a BLAST pairwise comparision of an *M. leprae* protein against the genome proteins of *A. fulgidus* and *A. aeolicus*. For each *M. leprae* protein, the BLAST score of the top protein hit in the *A. fulgidus* and *A. aeolicus* genome protein sets is recorded, and then the two BLAST scores are plotted against each other. Markers that lie between the diagonal (the dashed line in Figure 1) and the y-axis represent an *M. leprae* protein that is more similar to the top protein hit from *A. fulgidus* than the top protein hit from *A. aeolicus*. A marker that lies between the diagonal and the x-axis indicates that the protein found in *M. leprae* is more similar to an *A. aeolicus* protein than one found in *A. fulgidus*. Markers that sit on the diagonal represent *M. leprae* proteins that are equally similar to both species. Red markers delineate proteins involved in "Translation, ribosomal structure and biogenesis".

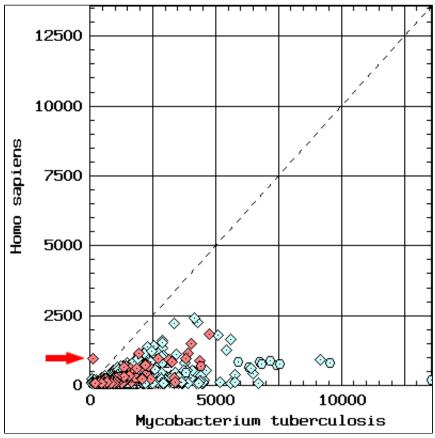


Figure 2

When *M. leprae* genome proteins are compared to genome proteins from *Homo sapiens* and *Mycobacterium tuberculosis*, a much different trend is observed (Figure 2). As would be expected, the vast majority of *M. leprae* proteins share greater similarity with proteins from *M. tuberculosis* than with proteins from *H. sapiens*. However, one protein, which turns out to be prolyl-tRNA synthetase (a product of the *pros* gene found out by clicking on the marker in a live query), shares greater similarity to the *H. sapiens* homolog. This match is marked by the red arrow in Figure 2. Such outlying points in TaxPlots may often be due to some kind of error, however in this case, the anomaly may be the result of a horizontal gene transfer between species. The comparison of domain architectures and multiple alignments has suggested that horizontal gene transfer has been a fairly common event in the evolution of tRNA synthetases of parasitic and symbiotic bacteria [1]. Futhermore, the genomic context of the *M. leprae pros* gene shows that it is displaced and inverted as compared to the *M.tuberculosis pros* gene. This supports the hypothesis that *M. leprae* acquired the *pros* gene by a recent horizontal transfer event [2].

Mycobacterium leprae genome (NC\_002677) was used as a query set to search a database of precomputed BLAST [3] genome protein alignments using standard parameters. The two-dimensional sequence alignments was constructed using TaxPlot.

- Wolf YI, Aravind L, Grishin NV, Koonin EV (1999) Evolution of aminoacyl-tRNA synthetases-analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res* 9, 689-710.
- Cole ST, et al. (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007-1011.
- Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Click this to close the window.